CAMUS DB
Development of Structural Database for Homology Search

Masayuki Kikuchi
mkikuchi@genes.nig.ac.jp
Sadahiko Misu
smisu@genes.nig.ac.jp
Tadashi Imanishi
timanish@genes.nig.ac.jp
Naruya Saitou
nsaitou@genes.nig.ac.jp

DNA Data Bank of Japan [DDBJ], National Institute of Genetics,
1111 Yata, Mishima, Shizuoka 411-8540, Japan

Keywords: CAMUS, DDBJ, homology search, compressed database, multiple alignment

1 Introduction

DDBJ/EMBL/GenBank International Nucleotide Sequence Database (INSD) is increasing with unexpected rate (doubling time is only slightly longer than one year), and computation time for homology search is becoming longer and longer. Because nucleotide sequences have their own nature to make copies through DNA replication, there are many similar, evolutionarily closely related nucleotide sequences in INSD. We therefore developed new system to cluster similar sequences and make compressed sequence database consisting of representative sequences of those clusters. Each cluster consists of highly homologous sequences, and they are presented as mulitply-aligned form. We DDBJ [1] call those combined database as CAMUS (Compressed database And MUltiple aligned Sequence database).

2 Method

Each division of INSD is treated separately. Because divisions were made according to major classification of organisms, such as VER (vertebrates) and PLN (plants), phylogenetically closely related sequences are expected to be already grouped in same division. We first sort all sequences in one division by nucleotide length, then sequences longer than 1.5kb were chopped into 1kb length fragments. Those short sequences were used as query sequence for BLAST [2] homology search (cutting point is more than 90% identity and more than 100bp). When one sequence is found to be homologous with query sequence, it will be eliminated from target database. By this procedure, evolutionarily highly homologous sequences were grouped into one, and set of representative sequences are called "Compressed Database". This database can be used as target database for any homology search. We so far implemented BLAST search for this secondary database.

Sequence groups with high homology are also created as byproduct of construction of Compressed Database. We devised a simple software to visualize multiple alignment by utilizing BLAST output directly. No further multiple alignment is necessary for those closely related sequences. This Multiple-aligned Sequence Database is also part of CAMUS DB, and this function is already in service in DDBJ homology search system [3].

3 Results

We used DDBJ Release 34, which includes 18 divisions. Because EST division has vast number of sequences from various organisms, we divided EST into 6 groups (human, C. elegans, Arabidopsis, mouse, rice, and others). Therefore, 23 sets were treated separately. After doing BLAST search, sequences were compressed up to 11% (PAT division), while GSS, STS, and EST (others) divisions remained uncompressed (88% , 90%, and 77%, respectively) in terms of number of entries. Compression ratios for other divisions were 13-55%. If we count number of total letters, most compression ratios are within 30-60% range. To check how homology search can speed up, we picked up M94937 (MHCDRB1 gene of chimpanzee). When PRI division of DDBJ database was used, 583 entries were retrieved, while only 85 representative sequences were retrieved when PRI compressed sequence data were used as target database. Because many MHC allele sequences exist in INSD, compression ratio in this case is quite low (85/583=15%).
Website URL of CAMUS DB is http://wolf.genes.nig.ac.jp/camus/.

4 Future Perspectives

This kind of structured database must be operated in near future, for nucleotide and amino acid sequence data are rapidly increasing. We are planning to construct similar compressed database based on our DAD (DDBJ Amino acid sequence Database). Current unit of sequence data submission (entry with unique accession number as its ID) is not convenient for homology search or other high speed data analysis. Those current system will be eventually considered only as archival purpose, and we should construct secondary database for sequence analyses.

Note. This article was written in "Scientific English" [4], which may be slightly different from common English.

References

[1] Website of DDBJ is http://www.ddbj.nig.ac.jp/.

[2] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215(3):403–410, 1990.

[3] Website is http://www.ddbj.nig.ac.jp/E-mail/homology.html .

[4] Saitou, N., Message from the Editor-in-Chief, *Anthropological Science*, 106(1): i-iv, 1998.